Variation in the local motion statistics of real-life optic flow scenes

Szonya Durant & Johannes M. Zanker

Department of Psychology, Royal Holloway University of London, Egham, Surrey, SW11 6HJ, UK

Szonya.durant@rhul.ac.uk

key words : natural scenes, expansion, 2DMD, modelling

Optic flow motion patterns can be a rich source of information about our own movement and about the structure of the environment we are moving in. We investigate the information available to the brain under real operating conditions by analysing video sequences generated by physically moving a camera through various typical human environments. We consider to what extent the motion signal maps generated by a biologically plausible, two-dimensional array of correlation-based motion detectors (2DMD) not only depend on egomotion, but also reflect the spatial setup of such environments. We analysed the local motion outputs by extracting the relative amounts of detected directions and comparing the spatial distribution of the motion signals to that of idealized optic flow. Using a simple template matching estimation technique, we are successfully able to extract the focus of expansion (FOE) and find relatively small errors that are distributed in characteristic patterns in different scenes. This shows that all types of scenes provide suitable motion information for extracting ego motion despite the substantial levels of noise affecting the motion signal distributions - attributed to the sparse nature of optic flow and the presence of camera jitter. However, there are large differences in the shape of the direction distributions between different types of scenes, in particular man-made office scenes are heavily dominated by directions in the cardinal axes, which is much less apparent in outdoors forest scenes. Further examination of motion magnitudes at different scales and the location of motion information in a scene revealed different patterns across different scene categories. This suggests that self-motion patterns are not only relevant for deducing heading direction and speed, but also provide a rich information source for scene structure and could be important for rapid formation of the gist of a scene under normal human locomotion.

1. Introduction

Optic flow, the characteristic pattern of motion signals that occur on the human retina when moving through an environment (Gibson, 1950), is an important way of gauging our own movement and gaining feedback on our direction and speed of heading. In the laboratory, human sensitivity to optic flow is usually tested using moving randomly positioned dots to in order to restrict input to motion information alone, thereby disambiguating motion from other cues such as object position and depth (Regan & Beverley, 1983; for reviews see Lappe, 2000; Warren, 2008). This however leads to a rather limited understanding of the kind of optic flow information available to humans in the natural environment and the ability to extract it. To tackle this question many past studies have taken the approach of using static natural images to generate artificial motion patterns calculated using the geometry to create appropriate motion signals. These generated motion fields have then been used to test motion models (e.g. Barron et al. 1994; Fleet & Langley, 1995) and description of motion statistics in natural scenes (Calow & Lappe, 2004; 2007), also combining recorded forward motion trajectories with static scenes to generate motion field statistics and using these to develop sophisticated prior probability based optic flow detector models (Roth & Black, 2007). A different approach has been to record the image sequence caused by movement through natural scenes and analyze the statistics of motion fields as calculated by low level motion detectors applied to the image sequences (Zanker & Zeil, 2005; Dakin et al. 2005). Both approaches present their own advantages - we take the latter approach for the closest approximation to genuine motion input experienced by humans and will argue that the motion model used in this work is a suitable approximation of early human motion detectors. This investigation offers a comprehensive analysis of motion direction and speed distributions caused by motion through natural scenes and also considers the extent to which motion signal distributions depend on the type of environment an observer is moving through. This work focuses on the output from local motion detectors as in Durant et al. (2011), but past work also exists where natural dynamic scenes have been analyzed for local and global motion content (Bartels et al, 2008).

Roth & Black (2007) considered artificially generated flow motion statistics over a large image database, too which they applied camera motion taken from a database of video clips (which could be walking or driving through a scene, but also around an object). Generating the motion field allowed them to asses the properties of the generated flow in comparison with the ground truth, one advantage of their method. They found that the generated image velocities were distributed in a characteristic pattern, peaking at very low magnitudes and dropping off with higher magnitudes and also found a great deal of horizontal motion with a smaller peak for vertical motion, however they state that much of this can be attributed to the distribution of camera motion used, which contained a great deal of horizontal motion.

Calow and Lappe (2007) took a similar approach based on earlier work by the same group (Calow et al. 2004). However, differing from our and Roth & Black's (2007) approach of using rigid camera motion they also reproduced typical eye movement data during navigation through scenes. They used a spherical model of the eye to reproduce retinal velocities. They considered the statistics of motion directions and speed relative to their location within the visual scene. They compared local motion directions with that of radial flow and saw systematic divergence and found a difference in speeds and the variability of speeds between lower and upper visual fields and with distance from the center of the visual field. Their extremely comprehensive and detailed work makes for interesting comparison with the current results using simple local motion detectors and real optic flow scenes of forward motion. As in Roth and Black (2007) this work generates and analyzes single frames of optic flow. This current work complements these analyses by considering optic flow generated in an environment over a larger time scale of several seconds and many frames.

Dakin et al. (2005) measured the distributions of different directions in movement through natural scenes recorded by walking and from a car and analysed using a simple motion energy local motion detector and found more motion energy in the cardinal motion directions rather than in the oblique directions, matching human performance patterns that show better motion discrimination around the cardinal directions - the so-called oblique effect. Their examples were all drawn from a

similar terrain of partly man-made park land. Advancing a camera on a gantry in a natural environment to record image sequences, Zanker & Zeil (2005) used a local motion correlation model that is functionally equivalent to some energy models and found that motion signals were often sparse and extremely noisy, rather different from the signals typically used as optic flow stimuli in psychophysical experiments.

Detecting self-motion is crucial in navigation for maintaining the speed as well as the direction of one's own movement in space, and a number of studies tried to identify the importance of the focus of expansion (FOE) in optic flow patterns for estimating the direction of heading (Lappe et al. 1999; Warren, et al., 1988). Detecting the FOE, and thus self-motion direction, is usually considered a problem of extracting global motion, but it is useful to consider first what is represented in local motion information, as basis of inferring global motion. The type, reliability, density, and distribution of this local information ultimately limits the global information that can be extracted. In this work we ask - how much information is present in local motion in natural scenes that can be used to extract optic flow? What can the statistics of local motion tell us about the structure of a scene? We investigate how optic flow generated local motion signals differ between environments. To assess local motion contents of a scene, we use the 2DMD motion correlation model as one of the simplest and biologically plausible models for local motion detection that has been used in a wide range of context to constrain expectations about the information available to biological systems (Zeil & Zanker, 1997, Zanker & Braddick, 1999, Zanker, 2001). This model performs a spatio-temporal correlation which provides a measure of the magnitude of the correlation - i.e. how well the image drives a particular motion detector unit at a given location - and we can use the relative magnitudes between horizontally and vertically arranged detectors to estimate the local direction of motion. The magnitude of the correlation can also give us an estimate of speed for some scenes, as natural images with their 1/f spatial structure (van der Schaaf & van Hateren, 1996) tend to be rather broadband, a case in which spatial-temporal correlation is closely related to speed (Meso & Zanker, 2009). The aim is for the results of the analysis to be independent on the model we have selected because most low-level motion models produce similar local outputs with broadband luminance-defined stimuli (Dakin et al. 2005, Durant & Johnston, 2009), but this will be further examined in the discussion.

Natural scenes are of particular interest, because the dynamic image sequences contain information not just about the direction of self motion, but also about the scene structure – about some aspects of shape, distance, size and the type of objects themselves (Wexler et al. 2001). In the local motion output these two things are not separate – the local shape of the object influences the shape of the flow, which a higher order motion mechanism or averaging process would need to correct for in order to extract global motion. Here we investigate how much this task will differ from scene to scene, by seeing how local motion varies and the systematic deviation that is caused by structure from an idealized expanding optic flow pattern.

The different scenes we compare are arranged loosely into three main categories of "office", "campus" and "forest" in ascending similarity to a "natural" environment that is not affected in its visual structure by human intervention. Analysis of natural scenes often involves making conclusions from statistics derived from large image databases (for a review see Simoncelli & Olshausen 2001). Our sample size is small but we wanted to ask how different scenes vary in their motion content, rather then combining them under the umbrella of natural scenes. To some extent these scenes represent on one hand everyday environments as experienced by modern humans, and on another hand the kind of natural environments that may have placed the constraints on our evolving visual system. This choice may give us a simple tool for having a glimpse at any differences between these two transformational pressures on the visual system exerted over different time scales.

2. Methods

Recording equipment and procedure

A Panasonic 3CCD miniDV tape camera was used to record AVI movies, with no image compression or motion stabilisation. The field of view was approximately 40° horizontally. The camera was attached to a trolley on wheels and sat 67cm above the

ground. The trolley was manually pushed along smooth wooden tracks with edges on to minimize side-to-side movement. The length of the track the trolley was pushed along was 200cm, which took approximately 10s, resulting in a speed of 0.2 ms^{-1}. Movies were recorded at 25 frames s^{-1}, resulting in about 250 frames per recording. The track could be moved to whichever location was needed. Movies were recorded indoors in office buildings ('office'), outdoors near to buildings around the campus ('campus') and outdoors in the middle of wooded area on campus ('forest'). Four different examples of each type of location were recorded and each scene was recorded at least 3 times (sometimes more), with the tracks in the same position.

Pre-processing

3 movie clips were chosen from each scene, judged by eye to have the least jerky, most straight-ahead motion. The AVI movies were split into individual black and white GIF images of 576×720 pixels, each frame having a maximum intensity value of 255 and a minimum value of zero. The first 50 frames were discarded from the start of the motion and then 150 frames used for analysis.

2DMD model settings

Image sequences were analyzed with a two-dimensional, correlation-type motion detector model (2DMD). This model has been used previously to analyze species-specific movement signals in crabs (Zeil & Zanker 1997), and to simulate a variety of psychophysical phenomena (Zanker 1997; Zanker & Braddick 1999; Zanker 2004) as well as investigate self-motion generated patterns in natural scenes (Zanker & Zeil, 2005). The basic building blocks of the 2DMD model are elementary motion detectors (EMDs) of the correlation type, which have been shown to describe the computational structure of biological motion detectors at least in insects (for review, see Reichardt 1987; Borst & Egelhaaf 1989). Under certain conditions correlation-type motion detectors become formally equivalent (see Hildreth & Koch 1987) to a variety of luminance-based motion detectors (Adelson & Bergen 1985; Van Santen & Sperling 1985; Watson & Ahumada 1985). We thus expect that the particular choice of model will not affect our main conclusions. Despite their differences in specific tuning

properties, these and other motion detector models (e.g., Torre & Poggio 1978; Srinivasan 1990; Johnston et al. 1999) are likely to generate similar distributions of motion signals, because they are based on local operations on spatial and temporal changes in luminance, although an important feature of the 2DMD is that its output is contrast dependent and hence it represents the very early stages of visual processing (Zanker et al. 1999).

In the simplest implementation, each EMD receives input from two points of a spatially filtered luminance pattern. The signals interact in a nonlinear way after some temporal filtering to provide a directionally selective signal. Difference of Gaussians (DOGs) are used as bandpass filters in the input lines, with excitatory center and inhibitory surround balanced as to exclude DC components from the input (cf. Srinivasan & Dvorak, 1980). The sampling distance between the two inputs is used as a fundamental spatial model parameter. The signal from one input line is multiplied with the temporally filtered signal from the other line, and two anti-symmetric units of this kind are subtracted from each other with equal weights, leading to a fully opponent EMD, which is highly directionally selective (Borst & Egelhaaf, 1989). The time constant of the first-order low pass filter was used as a fundamental temporal model parameter.

Unless stated otherwise the following default 2DMD model settings were used as follows - input filter gain: 4, EMD sampling width: 4 pixels (corresponds to approximately 0.25º a visual angle), Difference Of Gaussian (DoG) center width: 0.5 pixels (approximately 3’ of visual angle), EMD time constant: 6.0 frames (approx 250ms). These parameters were chosen after some piloting to find the largest output, the relatively long time constant is matched to the slow speed of movement forward. The model outputs motion correlations computed in the horizontal direction and in the vertical direction at each pixel for each frame. The first 6 frames of output are ignored in the analysis as the motion detector response peaks later in time after the initial input. The spatial filtering involved in calculating the motion process, leaves a boundary area at edges of each frame where motion calculations are invalid. The width of this area varies with the spatial scale of the filter and for a sampling width of 4 pixels is 36 pixels.

3. Results

Direction distributions

We begin by looking at the statistical distributions of directional responses in the different scenes. The direction of the motion responses is calculated at each pixel by finding $\tan^{-1}(hor/ver)$, where *hor* and *ver* are the horizontal and vertical motion outputs and can reach a maximum value of 1. The motion magnitude at each pixel is calculated as the magnitude of the vector formed of *hor* and *ver*, which equals $\sqrt{hor^2 + ver^2}$. Although we have only twelve scenes, for each scene we have many frames to inform our statistics. We begin by averaging over the output from all of the frames for each scene. We do not suggest that the brain averages over 144 frames as we do, as this would mean a temporal averaging in the range of 6s or so, which is not physiologically plausible. Rather, we are averaging to overcome any jerkiness in recording, arising from camera jitter, and to enhance signal-to-noise ratio. This will also emphasize the motion that is consistent over frames – i.e. the motion caused by forward motion, rather than any other motion due maybe to any intermittent motion of parts of the scene (such as branch movement). We average by averaging all the horizontal outputs and vertical outputs separately, keeping all the original measurements as far into the process as possible, and then extracting the directions from the two overall averages. This results in opposite horizontal and vertical directions cancelling each other out, which would be the side-to-side or up and-down motion of camera shakiness. It also avoids on each frame very small motion magnitudes resulting in spurious motion directions and allows small consistent motion direction to add up.

First we consider the visualized 2D directional output (Figure 1). We can consider the direction only output (parts (ii)), which shows the direction (as indicated by the color wheel) for every pixel where motion was measured above some minimal cut-off point (below threshold is shown as white), or we can consider the direction and magnitude output, where the directions are scaled by the motion magnitude (parts (iii)); here color indicates direction and saturation indicates motion magnitude. Just by qualitatively considering these example outputs we can see differences between the

scenes - the less artificial scenes appear to contain more continuous motion, and there are greater left/right signals in the artificial scenes that contain more edges and sparser, but less noisy motion information. The streaks are caused by averaging over time and show that over this length of time, the change in the motion pattern captures some of the optic flow structure. All images in these examples appear to exhibit a motion pattern consistent with optic flow, namely that of expansion. There does not appear to be a great deal of variation in the maximum magnitude, more the different locations of motion in the scenes.

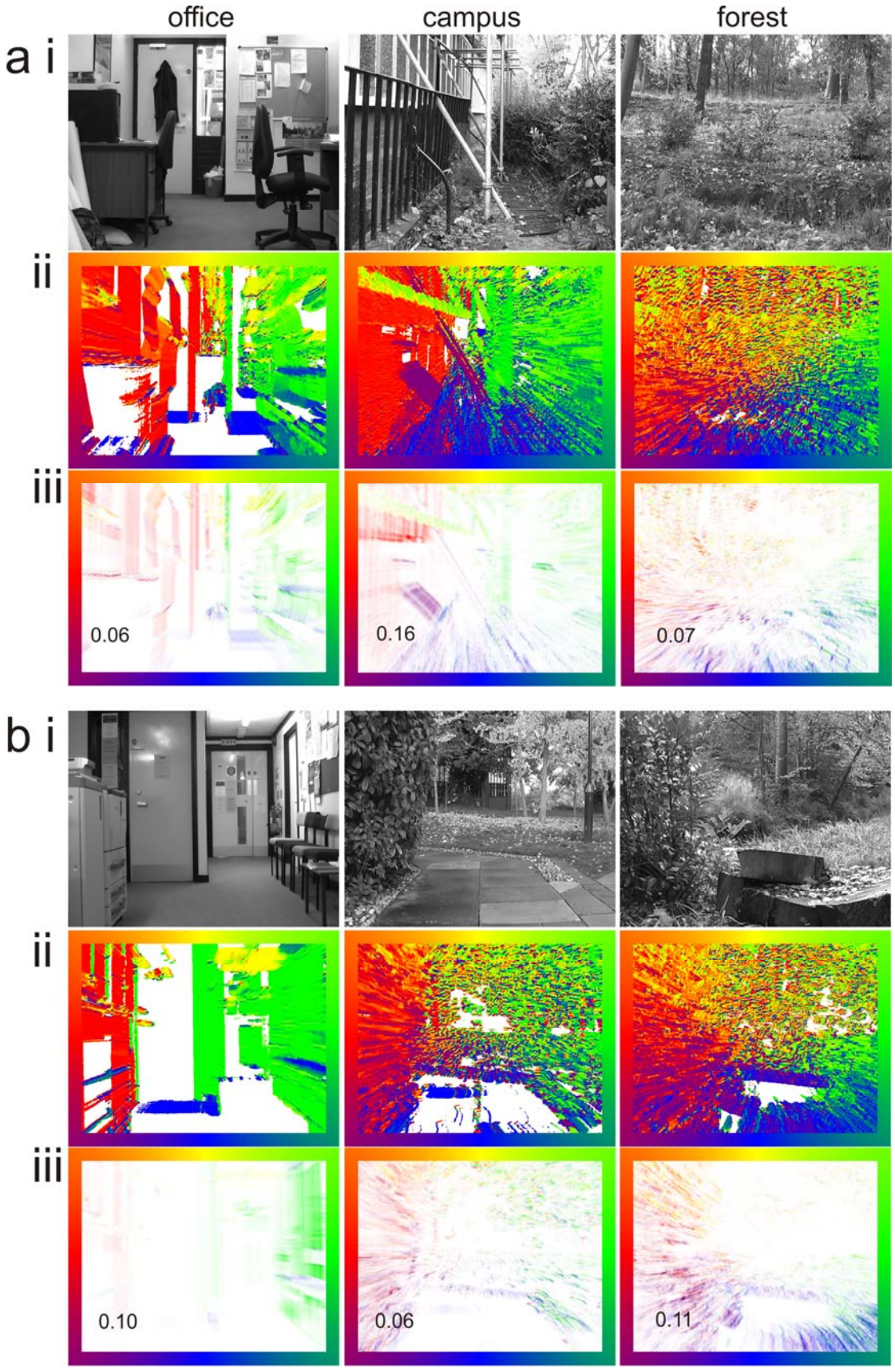
office
campus
forest
a i
ii
iii
0.06
0.16
0.07
b i
ii
iii
0.10
0.06
0.11

Figure 1. Illustration of averaged direction outputs. (a (i), b(i)) A single frame is shown from each of the 6 example clips. (a (ii), b(ii)) Direction only outputs as calculated from the separately averaged *hor* and *ver* outputs over all the 144 frames (thresholded to zero for motion magnitudes less than 0.1% of the maximum output, zero values are white, the color wheel shows direction). (a (iii), b(iii)) direction and magnitude outputs of the same averages as in (ii), with color saturation representing the magnitude, again white represents zero values. The images have been normalized to their own maximum, for which the value is given in the bottom left hand corner.

It should be noted that these flow fields merely show how much information is available in principle to in these scenes to extract optic flow, but due to the long averaging times are most likely to be an overestimation of what is available to the visual system. In Figure 2 we demonstrate just how sparse and noisy the information is if we average over 10 frames and why we chose to use the long averaging times to demonstrate what information may be available if jerkiness of motion is removed and signals are aggregated. Note the maximum values are lower when averaging over more frames and this is because the opposite side-to-side motion cancels out, for instance in the forest scene on the right in Figure 2, there is leftward bias over these 10 frames. We will analyse the aspects of the available motion information quantitatively by averaging over all 144 frames from now on.

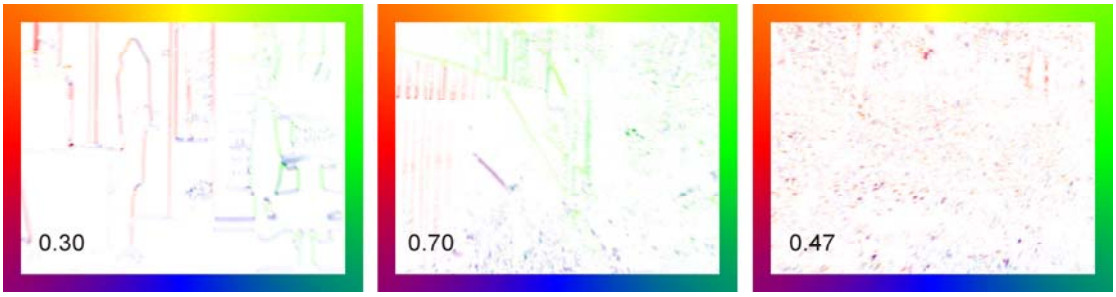

Figure 2. Averaged direction outputs for 10 frames for the clips in Figure 1(a), with color saturation representing the magnitude, again white represents zero values (thresholded to zero for motion magnitudes less than 0.1% of the maximum output), the color wheel shows direction. The images have been normalized to their own maximum, for which the value is given in the bottom left hand corner.

We now quantify the motion intensity present for each direction of motion for the different types of scenes. We do this by first finding the average horizontal and vertical motion magnitudes over all of the frames as above. This allows all motion values to contribute, without the problem of ascribing meaningless directions to very small motion magnitudes. We then label each pixel by its direction using the averaged *hor* and *ver* outputs as before, to the nearest degree. We then sum all the *hor* and *ver* outputs of pixels (still keeping them separate, to maintain reliance on the original data and avoid rounding errors for as long as possible) that have direction of 0º (rightward) and so on for each angle to 259º, to provide a summed *hor* and *ver* magnitude at each direction. We then find the magnitude of the vector formed from the summed *hor* and *ver* values. This is done for all pixels within a circle of radius 252 pixels from the center (within the cut-off area boundary around the image edges left by the motion filtering process). This results in a representation that shows the amount of motion signals in each direction in the averaged sequence.

First we compare recordings of the same scene over three repetitions to assure ourselves that variation within a scene is small enough to make comparison between the different scenes meaningful. Although there is some variation in magnitude, the overall pattern of direction distributions remains the same within a scene (see Figure 3a, b). When we average all the examples within categories (see Figure 3b), we see distinct patterns emerging although there is huge variability between the different examples. The different peaks and skews in different images are caused by the different arrangements of orientations in the scenes and where the areas of high contrast are to be found. Some bias may also occur if the focus of expansion was not central to the scene (if the camera was not pointing straight ahead). We average the normalized histograms (divided by the number of pixels, which is the same for all scenes) in an effort to get the average direction pattern regardless of which scene generated larger motion correlation magnitudes. The forest scenes have directions distributed fairly evenly amongst all the directions, whilst in the office scenes virtually all the motion energy lies along the cardinal directions, with a large bias towards the horizontal directions. The campus scenes lie somewhere between the two, as they are more peaked than the forest scenes, but more broadly distributed than the office scenes. They appear to show less motion upward, which may be due to more sky being visible in these scenes. Reassuringly all scenes look different to the output

produced by noise (images of half back and white pixels, randomly generated one each frame, also for 144 frames), suggesting some kind of regularity in all movements through real-world scenes, despite the apparently noisy output.

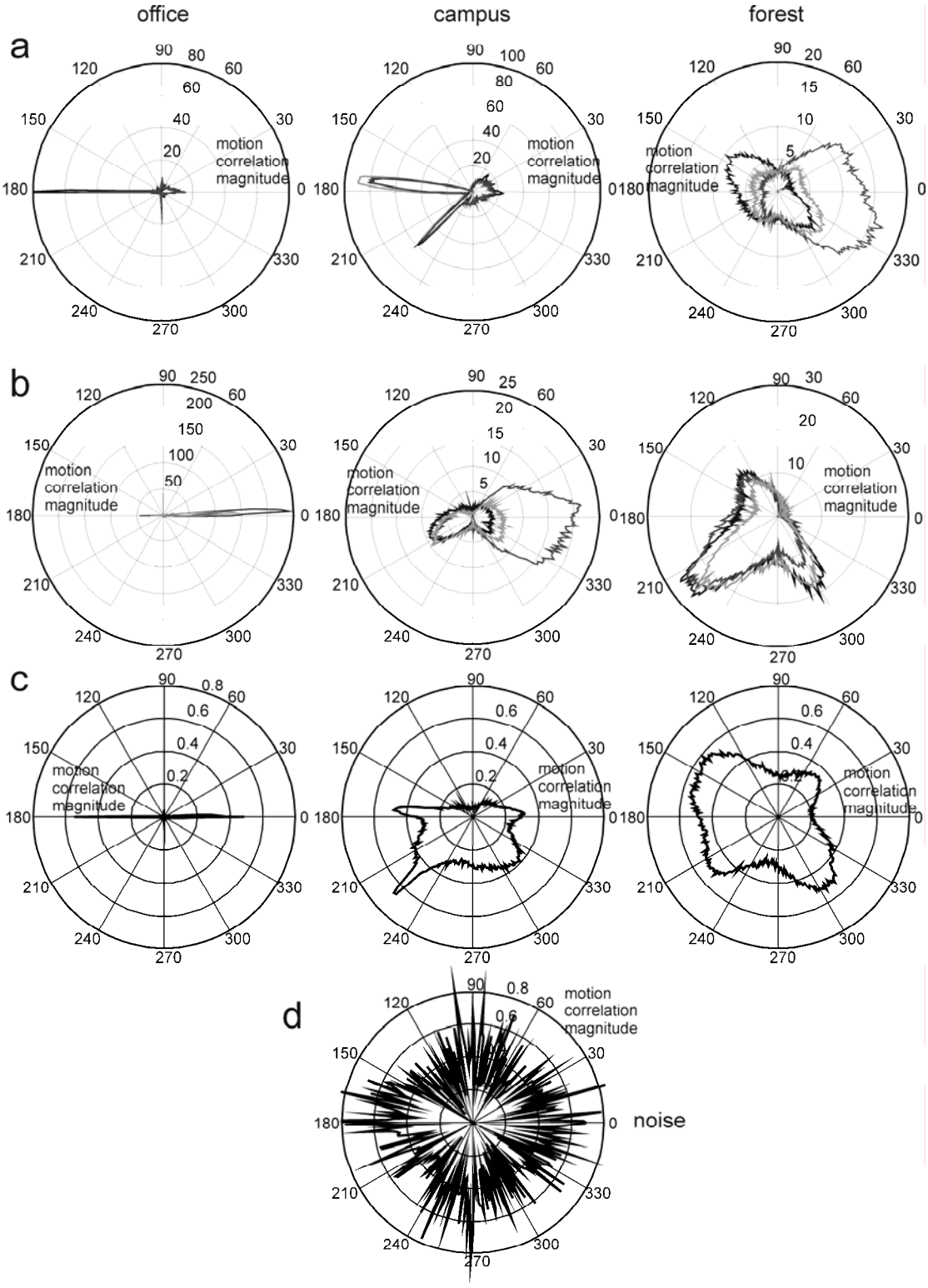

Figure 3. The distributions of motion signal directions calculated from the averages of the *hor* and the *ver* outputs over 144 frames. The direction of motion is plotted as the polar angle and the radius is the average magnitude of motion correlation at that angle of direction. (a) Magnitudes derived from the example clips from Figure 1a, with the results from the different instances of filming shown in the three different shades of gray. (b) Magnitudes derived from the example clips from Figure 1b. (c) The average of the normalized histograms from all four clips from each of the category of scenes. (d) The normalized distribution for a sequence formed of 144 randomly distributed (independent for each frame) pixel black and white noise dots, which generate random motion directions.

Furthermore we average the raw histograms (Figure 4), rather than the normalized histograms (as in Figure 3), over each of the cardinal axes, in quadrants, averaging angles 45º to 135º, 136º to 225º, 225º to 315º and 316º to 44º, on a scale of -45 º to +45 º relative to the cardinal axes (90 º, 180 º, 270 º, 360 º), to highlight the distributions relative to the cardinal directions. We can see even more clearly that indoor built environments (office) show a huge bias to the cardinal directions not seen in natural outdoor settings (forest). In the campus based clips, which are somewhere between the two, although the histograms are of a more similar shape to the forest scenes, there is a slight bias at the cardinal (although less sharp than in the indoor scenes and not exactly centered on the cardinals). In these scenes it is possible that the trolley was not exactly horizontal and yet, there would be strong vertical and horizontal cues in the scene, which could cause the small peak at off-cardinal direction). There is a slight increase in the amount of energy found at the oblique directions also - this seems to be caused more high contrast edges in these scenes that are at 45º, such as roofs and props against walls. It is interesting to note, that although both the office and campus scenes contain the edges of paths, floors etc., which appear as oblique angles in a 2D representation, these do not contribute to oblique direction signals as they are parallel to the direction of motion and thus no motion is detected due to the aperture problem that which states that an edge moving within a small aperture in the same direction as its own orientation will not produce a motion signal. In general over all angles away from the cardinals (i.e. towards the diagonals) the forest scenes have the highest response.

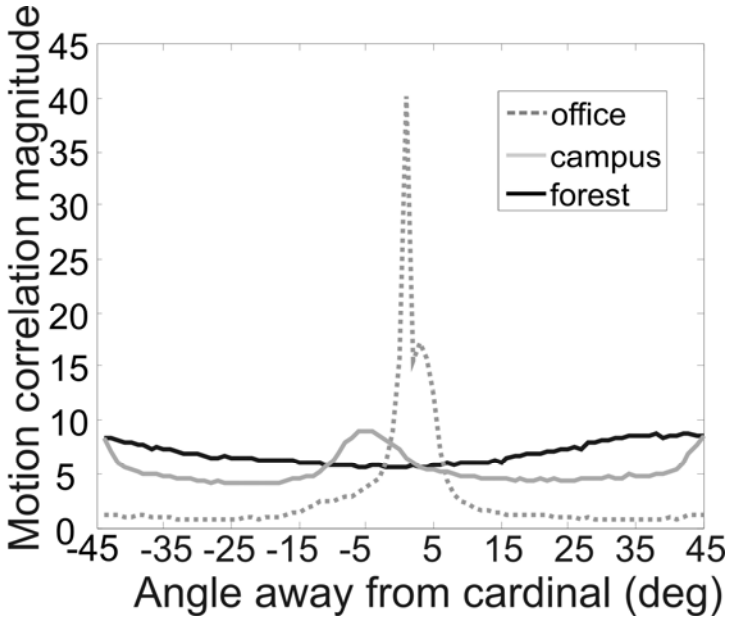

Figure 4. Histograms of intensity of motion at each direction averaged over direction quadrants, centered on the cardinals. Taking the histogram of 45º to 135º , with midpoint 90º and defining it as -45º to 45º, with midpoint 0º, the cardinal direction. Similarly taking 136º to 225º with midpoint 180º and redefining it as -45º to 45º, with midpoint 0º, and so on for 225º to 315º and 316º to 44º. Now each of these histograms defined on a scale of -45º to 45º can be averaged together. This way we see the distribution relative to the cardinal angle - any cardinal angle - on average.

Optic flow

The relevant optic flow pattern in our scenes is caused by forward motion i.e. a radially expanding pattern. Therefore, although optic flow in general can refer to any pattern caused by our own movement, we use the term optic flow from now on to specifically refer to radially expanding flow. To assess the quality of the forward motion information contained in each of the different scenes, we compare the distribution of motion outputs from the sequences with an idealized forward motion pattern. In an idealized expanding optic flow pattern, with the FOE at the center of the image if we trace a radial line from the center of the image (and hence the center of

expansion) to the edge of the image and sum all the *hor* and *ver* values along this line and then take the direction of the vector formed by the sums then this will be the same angle as the angle of the radial line originating in the image center. For example if we trace a line from the center straight out to the right edge of the image, the direction of the vector formed by the summed *hor* and *ver* values would be 0º (i.e. no vertical and only positive horizontal motion components), which is the angle of this line in the image. In other words, a FOE in the center of the image leading to radial expansion motion is always streaming away from the centers into the periphery. Calculating the absolute angular difference (avoiding wrap-around issues) between the motion direction output along a given radial direction from the image center and the angle of this radial direction, we derive an error measure for the FOE estimate from the motion signal distribution.

However, although an effort was made to fix the camera on to the trolley pointing straight ahead and to avoid any sideways motion of the trolley, some of the scenes have centers of expansion that are not central to the image. This is shown in the difference between recordings of the forest scene in Figure 3b and c. In one of the recordings there was a larger bias towards rightward direction, implying a *leftward* motion component of the camera center (see Figure 5a) and vice versa in another recording (see Figure 5b). In order to extract the direction of heading from these scenes we need a method to find the center of expansion. Because of the nature of the ideal flow being symmetric horizontally and vertically we could use this for our search strategy for matching the output to the ideal flow. We first choose in turn each of the horizontal positions, with the vertical position fixed in the middle. Each point was used as hypothetical center of expansion and the point that resulted in minimum error as described above was chosen as the horizontal position. We then fixed this horizontal position and varied the vertical position to find the minimum error. This differs somewhat from the method used by Zanker & Zeil (2006), who simply averaged directions across each horizontal position in the frame and fitted a simple left/right direction square wave function. The horizontal position that provided the best fit was used, which would give you an estimate for the FOE in fronto-parallel coordinates. Our method uses the matching of a more precise 2D optic flow template that also allows for estimation of the height of the FOE. Checking by eye appeared to show that this method was fairly accurate at detecting the center of expansion in most

cases. Although we do not know the true heading direction, the averaged direction output gives us a sense that is confirmed by the automated process. It should be noted that of course heading estimation is much more complicated in natural situations that involve rotations of the eye, head or path (Beverley & Regan, 1982), but in this simple case where the heading is mostly straight ahead on a straight path with no eye movements, this simple estimation method will suffice.

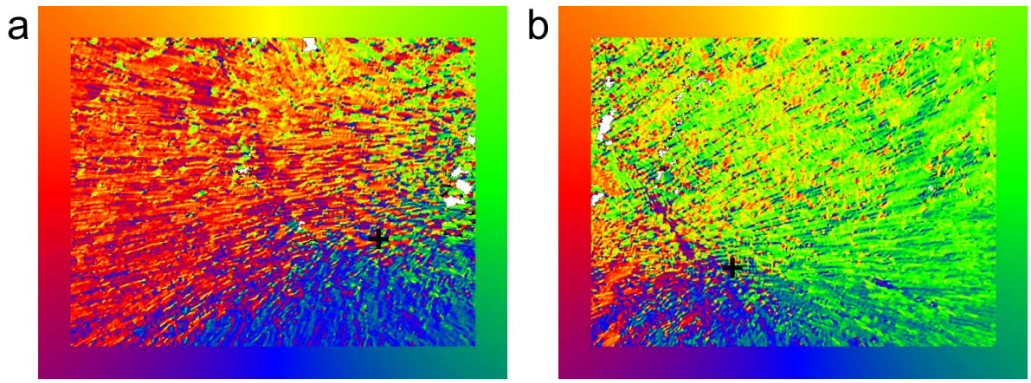

Figure 5. Examples of average motion responses (thresholded and scaled to maximum, see Figure 3b and e, white areas have motion magnitudes below 0.1% of maximum possible output) over 144 frames for sequences from the same scene. The FOE as automatically located by our method described in the text (shown by a black cross).

Figure 6 shows plots of the error as a function of radial angle from the calculated FOE, where each example clip from a scene is represented in different shade of gray. We find the smallest error in the office scenes (Figure 6a), with systematic errors around the oblique directions and more errors along the vertical than the horizontal. Although these scenes are dominated by horizontal signals, these do not appear to cause great error in the vertical direction, there appear to be enough horizontal orientations in the correct parts if the scene to cause the appropriate vertical motion (as mentioned above a contour can only cause local motion orthogonal to itself, not parallel). In the scene represented by the medium gray dots in Figure 6a, the bottom left part of the scene does not contain much structure and hence not much useful motion information. In the campus scenes the largest errors occur around the upper

half of the image, where in one case there is sky, and in another far away buildings, and therefore less motion information causes error in these cases. This is similar to the overall findings from Calow & Lappe (2007), where on average they found that the difference from radial motion was greatest in the upper visual field. If there is no motion to speak of, random noise just causes errors (e.g. blue dots in Figure 5b). In the forest scenes optic flow mostly matches well with expansion, with no clear systematic errors, apart from occasional noisy patches in the scene shown by the red dots. This also illustrates the additional information that can be seen by not averaging all types of scenes together and retaining categories.

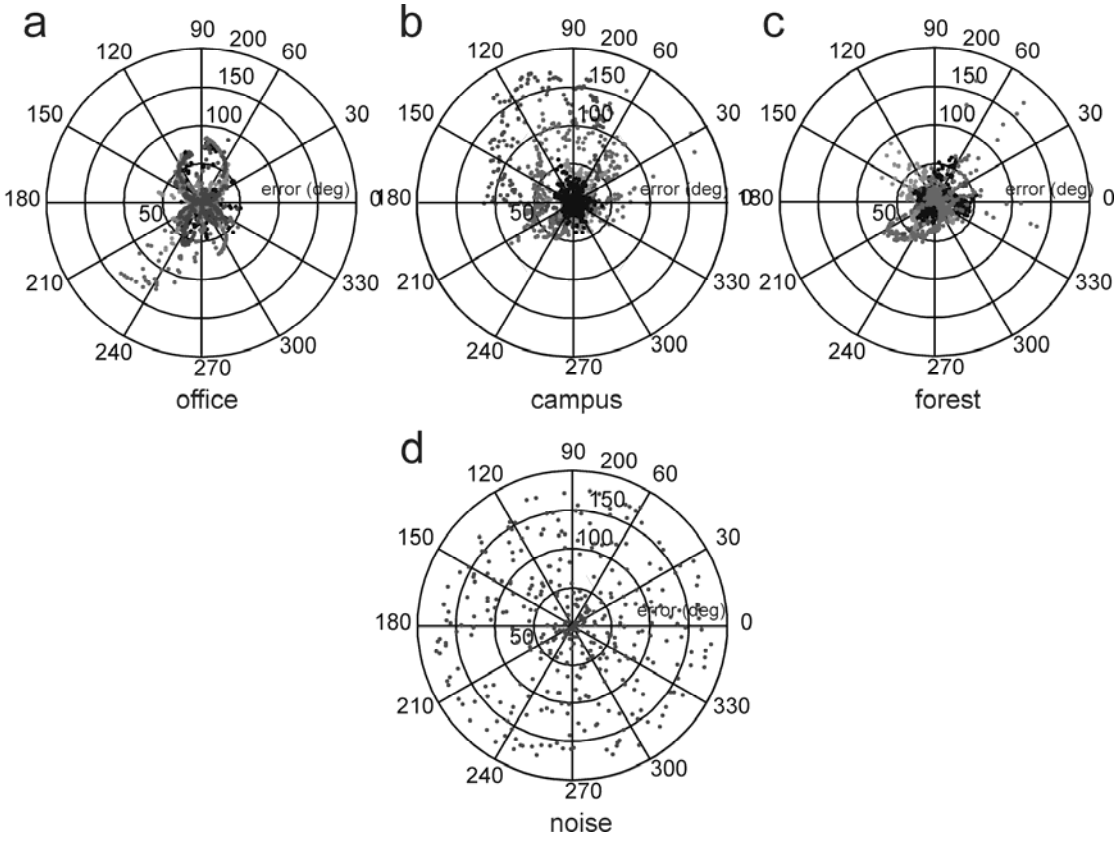

Figure 6.The absolute error calculated between the angle in the image (from the center of expansion) and the average direction of motion calculated along that angle calculated for the four examples of each type of scene (a-c) and for a random dynamic noise pattern (d). The different shades of gray dots represent the different examples of each scene. The maximum possible error is 180º in direction. There are more errors in the forest (a) and campus (b) scenes,

especially around the upper half of the scene where there are often less motion signals.

Magnitude distributions

We also calculated the magnitude of the motion signal at each pixel for each frame and binned all these in 256 bins from zero response to the maximum of the response range. This generates a distribution of the motion energy of the response maps that shows that most of the responses are small (Figure 7a). The distributions have a characteristic shape similar to that found by Roth & Black (2007), when they artificially generated motion sequences. Although they kept horizontal and vertical velocity separate and we measure motion correlation instead of velocity, as mentioned above the two are related, and it is interesting that the real-life motion measured here in terms of local motion correlation generates similar patterns. This also agrees with the work of Calow & Lappe (2004) who also found a peak at low speeds. We also summed all the motion magnitudes over all the pixels and all the frames within each category to make an overall coarse comparison of the amount of local motion (Figure 6b) and find more motion as the scenes become less artificial. This can be explained by the sparseness of the office scenes and also the increasing noisiness of the campus and forest scenes. We also considered whether the amount of motion in the scene helps the extraction of optic flow (Figure 7c) and found no correlation ($r=-0.20$, $p=0.523$) between the amount of motion and the average angular error away from the ideal optic flow for each movie clip. This observation supports the view that additional motion signals in outdoor scenes may be caused by other sources of motion such as the motion of leaves in the wind and noisiness in the scene, whereas clearly in areas with no motion, such as the sky, optic flow extraction is not possible.

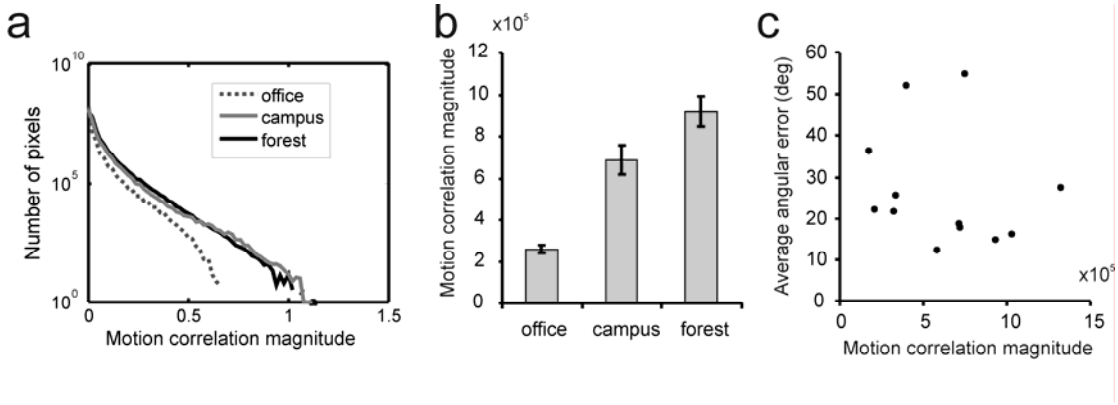

Figure 7. Measuring the overall magnitude of motion correlation in different scene categories. (a) Frequency histograms of motion magnitudes, plotted for each pixel in each frame of each sequence, for three types of scenes (log scale y axis). Maximum horizontal/vertical motion output at each pixel is 1, so maximum possible magnitude is $\sqrt{2} = 1.41$. (b) Areas under the curves summed and plotted for ach type of sequence. Error bars are standard error over the four clips in each category. (c) Optic flow error (summed over all angles (shown in Figure 6) in the averaged sequence output) plotted against average motion correlation magnitude.

Varying spatial scale

So far all our results are calculated at a single spatial (and temporal) scale of the elementary motion detector. Motion detection happens at several spatial scales and it could be helpful for global motion extraction to combine the information across these scales. It is therefore important to consider what kind of information is available at different scales and whether the patterns we have observed so far change with spatial scale. By considering the two example scenes in Figure 8 (which are the same as the forest and office scenes in Figure 1a), we can see that the general pattern remains the same, although the amounts of motion at different spatial frequencies varies.

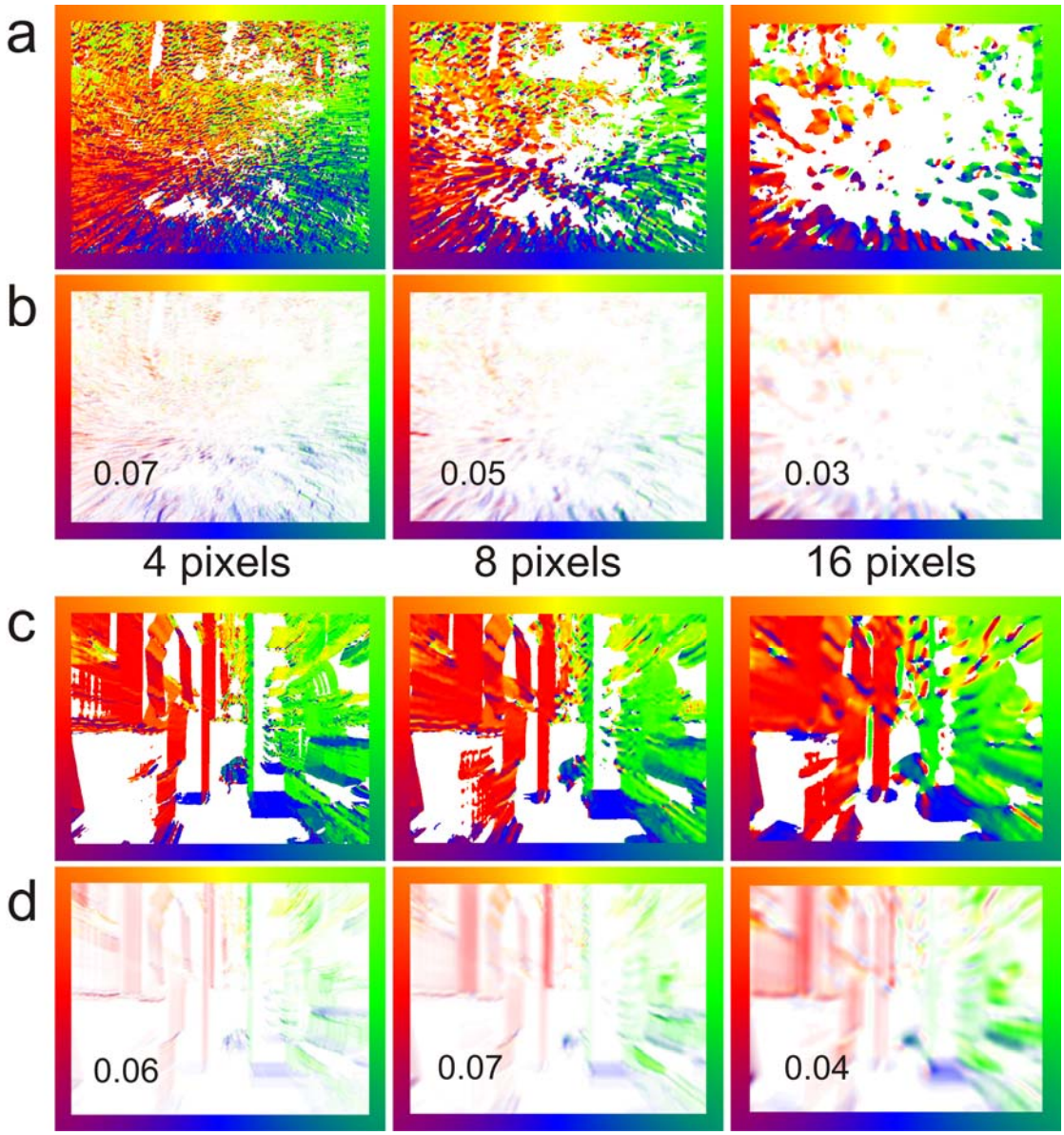

Figure 8. Motion outputs for forest (a, b) and office scene (c, d) from Figure 1a. Shown for different spatial scales (sampling width of motion detector given in pixels for 4 different columns) showing direction alone (a, c) and combined direction and magnitude (b, d), scaled to the maximum motion magnitude, for which the value is given bottom left corner.

To further consider the magnitude of the motion outputs as a function of spatial scale, we summed motion correlation magnitude over all pixels in each of the sequences and compared across scales. At larger scales the larger sampling distance means that motion magnitudes can be calculated for less of the scene, with larger boundary regions where motion can not be calculated, (shown by the wider boundary areas in

Figure 8b and d). All magnitudes are summed over the valid area at the largest scale, with a boundary of 42 pixels. In Figure 9a we see that for all types of scenes there is a decrease in the amount of motion correlation as sampling size increases. A one way repeated measures (i.e. controlling for variability introduced by different mean amounts of motion in each clip) ANOVA on our samples confirms this is significant for each type of scene (office: $F_{2,6} = 6.35$, $p<0.05$; campus: $F_{2,6} = 16.46$, $p<0.005$; forest: $F_{2,6} = 20.88$, $p<0.005$). In contrast to the other scenes the office scenes only show a reduction in the summed motion magnitude at the largest scale, showing that the amount of motion correlation does not follow the same pattern across scales in different categories. In Figure 9(b) we compare how the ability to estimate the FOE varies as a function of spatial scale and find that in this case error produced by the information available in forest and campus scenes increases at larger scales, (campus: $F_{2,6} = 6.49$, $p < 0.05$; forest: $F_{2,6} = 16.82$, $p < 0.005$) but the scale has no effect on the amount of error in office scenes (office: $F_{2,6} = 1.86$, $p = 0.26$). The relationship between motion magnitudes per se does not necessarily predict the ability to extract optic flow, as office scenes at small spatial scales contain much less motion signal than the other scenes, yet this does not seem to affect the ability to extract optic flow. This suggests that a large proportion of the motion correlation signals in the forest scene arise from noise that is not informative of the FOE. The fact that the largest motion signals are not necessarily the most informative about direction is reflected to some extent in the findings from Calow & Lappe (2007), who found minimal spatial correlation between direction and speed estimates and found different patterns of distribution for direction and speed signals.

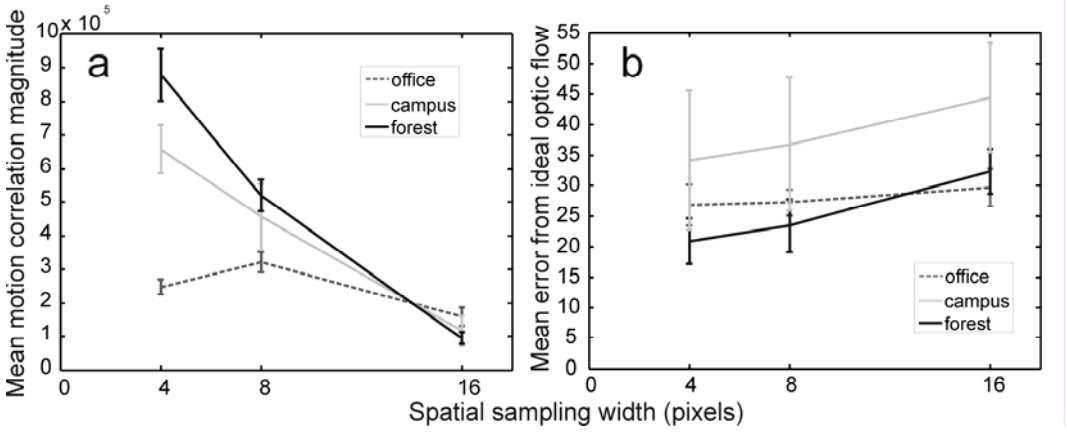

Figure 9. (a) Motion magnitudes summed over each pixel in each frame of each sequence (within the boundary limit constrained by the cut-off area under sampling distance 16, see text). Sampling width corresponds to the scale at which motion is detected. Motion output magnitude shown, where maximum possible output is horizontal or vertical output ate ach pixel is 1. (b) Summed angular error relative to ideal optic flow summed over all angles calculated based on the averaged horizontal and vertical outputs (within the valid region left by the largest spatial sampling width).

Eccentricity

The final aspect to consider is where the motion information can be found in the visual scenes, and how this changes over different spatial scales. In order to do this, we divide each scene up into a circle immediately around the FOE as calculated above (assuming this is what would be focussed on the fovea) and into two rings further away out from the FOE. We chose three equally spaced radius distances from the FOE. Magnitudes were averaged over the number of pixels that lay within each area within the image (as sometimes the FOE was not central, some parts of the outer rings were not contained within the images, making for unequal areas). The average over the number of frames was used as a more meaningful measure in this case.

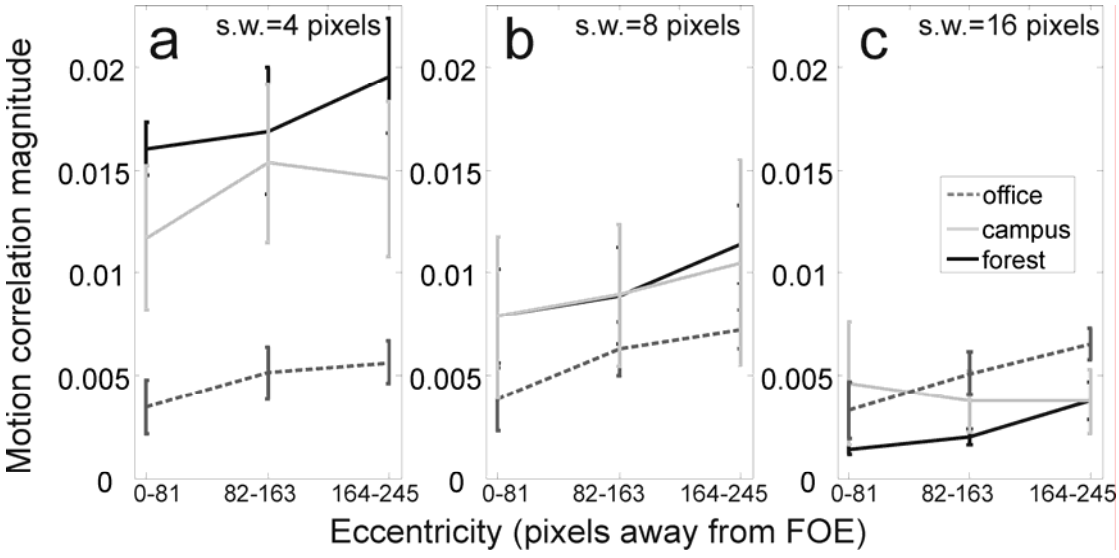

Figure 10. Average motion correlation magnitudes (ordinate) at different eccentricities (abscissa) with increasing sampling width (s.w. 4, 8, 16 in a-c), for three different types of scenes (see legend). Error bars indicate standard error over the 4 scenes in each category.

We again find a decrease in the amount of motion correlation with increased sampling width, whilst generally the same general trend is maintained that slightly larger motion magnitudes are found at higher eccentricities. This is to be expected as there are larger displacements caused in the periphery by a typical optic flow pattern, whereas there is not much motion typically at the FOE. The pattern doesn't change much with spatial scale, just the overall amount of motion. We find that for all types of scenes the amount of information increases as we move away from the FOE, and this holds true across all scales, apart from in the campus scenes. For the forest scenes there is significant effect of eccentricity only at the largest scale ($F_{2,6} = 5.92$, $p<0.05$), whereas for the office scenes there as a significant effect of eccentricity at all scales (s.w.4: $F_{2,6} = 5.97$, $p < 0.05$; s.w.8: $F_{2,6} = 7.84$, $p < 0.05$; s.w.4: $F_{2,6} = 13.40$, $p < 0.01$). As the office scenes are broadband we expect greater velocity to correlate with greater motion correlation (Meso & Zanker, 2009) and optic flow causes greater displacements in the periphery - and this is what we observe. Again, this is a pattern we observe in the work of Calow & Lappe (2007) where in the overall distributions velocity increases towards the periphery of the visual field. This effect may be reduced in the other scenes due to the distance of objects in the periphery, yet it is interesting that in forest scenes it emerges most at the large spatial scales, which coincides with where the most error was to be found in the FOE calculation, possible because a great deal of the periphery is not captured in the scene.

4. Discussion

Much has been made in the past of the properties of natural scenes as the training set for humans, with strong explanatory power for the properties of the visual system. This explanation can occur at two levels: that of a visual system shaped by evolution in such a way that it is well adapted by birth to the environment and also a visual system that is plastic be wired over a lifetime in such a way as to be adapted to

incoming stimuli. Natural images are often considered a special class of stimulus in themselves, with statistical properties that hold a great deal of explanatory power such as the 1/f power spectrum observed (van der Schaaf & van Hateren, 1996) and the sharp differences between the amount of oblique and cardinal orientations (Coppola et al., 1998). These properties have been observed in stationary stimuli and can help us understand a great deal about human psychophysical responses to visual stimuli. Furthering this work on stationary images we have analysed the statistics in dynamic stimuli, which to date has been carried relatively rarely in realistic dynamic scenes of motion through an environment (Dakin et al. 2005). Some work has looked at recreating the dynamics of movie scenes and is important for comparison as in these ground truth is known, unlike in our stimuli (Roth & Black, 2007; Calow & Lappe, 2007).

In contrast with these studies of Dakin et al. (2005) and Calow & Lappe (2007) we find that the bias towards cardinal directions for motion does not necessarily hold true in all types of scenes. Not surprisingly, it is a definitive feature in the more artificial indoor scenes, which contain many horizontal and vertical orientations. The sparse nature of these scenes means that local motion measurements are more subject to the aperture problem, leading to only the detection of motion orthogonal to these high contrast edges. However it is not at all present in our collection of outdoor scenes, despite the presence of vertical structure such as tree trunks, the cardinal motion appears to be drowned out by the other motion directions present. In this case the edges are often lower contrast and are also surrounded by dense texture that would reduce the aperture problem, leading to a signal less biased in the orthogonal direction to the contour. Of course, in natural scenes not all this motion is due to self-motion, but also the movement of branches in the wind for example, something the reconstructed motion from scenes in the Roth and Black (2007) study would not contain. This work serves to highlight these differences, but it is for further work to investigate in the detail the cause of these.

We did replicate from previous studies the shape of the motion magnitude distribution (Roth & Black, 2007; Calow & Lappe, 2007). We find the same distinctive shape, showing that this is true also for real life motion and that our estimates of local motion are similar despite using the 2DMD model first rather than ground truth motion. This

supports other findings suggesting that if we consider the motion system as operating within a Bayesian framework it may assume a prior probability distribution centered on zero velocity (Weiss et al. 2002). The use of the term ‘prior’ indicates the shaping of the visual system over the lifetime, whereby the experienced distribution of speeds influences subsequent perception of speed.

By successfully using the motion directions to find the FOE in most cases, we were able to quantify the amount of error in matching to an optic flow template. We find this error mostly to be low, demonstrating that local motion contains the information needed to extract optic flow (although due to long the averaging times used here it is not clear how the visual system is able to extract it), even though it is sparse and noisy. This sparseness of motion in natural scenes found here agrees with previous findings (Zanker & Zeil, 2005). We find this even as our camera physically moves through the natural scene, rather than just advancing on a short gantry. However, there are characteristic deviations from flow, shown clearly in the motion direction histograms, away from a perfectly circular even representation of each direction. A great deal of local motion is due to the image structure (anything that is different between the scenes, as self-motion was the same in each scene), so gives clues to both self-motion and the contents of the scene. Calow & Lappe (2007) considered which areas of the visual scene provided more information in terms of structure and found the lower region of the visual field less informative. In their work they consider structure in terms of depth information, whereas in the work here we refer to the 2D contour information, which will be characteristic for different types of scene. We found large differences between the different types of scenes, with clear characteristics of different kind of scenes becoming apparent even over these few examples. To simply class these together as part of a set of ‘natural images’, could be misleading. These differences may provide a way of teasing apart evolution versus development. For instance, this work ties in with well-known past physiological studies on cats reared in artificial environments, where biases in environment were reflected in the cat physiology (Wiesel, 1982). Similarly if indeed oblique versus cardinal differences exist in human neural make-up as reflected in processing ability this may be due more to the built environment surrounding humans now then the kind of natural environment that may have driven human visual evolution.

This work is also of relevance to the literature on 'gist' perception. Past research has investigated our ability to rapidly categorize different kinds of scenes (Friedman, 1979; Rousselet et al., 2005). These studies have shown that humans are able to form a quick overall impression of a scene without necessary perceiving it in detail to make a categorization choice. The current results suggests that local motion information could also provide considerable cues to the type of scene we are moving through and so could be important for gist perception, which in real life situations must be performed during locomotion through an environment and be required for rapid adjustment of motor outputs.

At different scales our findings remain the same up to a point, but high spatial frequencies play a greater role in the motion patterns of some scenes than others, and in general, where there is more motion more optic flow can be extracted, although this is not a very strict relationship. Relative differences in motion magnitudes between images change as the spatial scale changes. As our images are largely broadband, the motion correlation can be expected to be similar at all scales, however it seems we would need to combine over many spatial scales to get a comprehensive picture as suggested by Meso et al. 2009. Although optic flow is a global percept, it seems in order to extract it, at first we may need to take into account high spatial frequency local information.

When examining the spatial layout of the motion signals we found that more peripheral areas contained more of the motion information, which is interesting because this might be an additional way in which the visual system is matched to commonly experienced stimulus properties as contrary to other visual attributes sensitivity to motion does not decrease in peripheral vision (McKee & Nakayama, 1984; Lappin et al. 2009). This would tie in with the efficient coding scheme proposed by Calow and Lappe (2008). Of course not all fixations during self motion are at the FOE, especially when negotiating obstacles (Hollands et al.,1995), but when moving straight ahead in an obstacle free environment is seems that most of fixations are near to the FOE (Wilkie & Wann, 2003). In this work we have not been able to comment on the very important factor of eye movements in general as the camera angle is fixed, and it should be noted that a similar pattern of motion magnitudes may be caused by smooth pursuit movements. Mobile eye tracking technology should be

able to open up this whole area more in the future to be able to measure more realistic input.

We now arrive at a brief consideration of the motion model used. It is a simple correlation model and we claim that these results are not specific to artefacts in this model. We showed that with a black and white noise input there are no great artefacts and we have seen that the scale we chose didn't matter greatly to the results. This model differs from some other models such as gradient based approaches (Horn & Schunk, 1982; Johnston et al., 1999) in that it doesn't extract velocity and it is contrast dependent, so the results need to be understood in this context. A large part of the spatial variation will be due not only to more motion in that area, but to higher contrast, but this is indeed what local motion signals will initially reflect in the human visual system. These are the initial limitations that we are describing and at later levels velocity and contrast independent motion signals as well as over all optic flow directions may be extracted, but they will all be dependent on the motion information that is available at this level. Therefore our conclusions aim to generalize to the concept of local motion in general, not just in terms of the 2DMD model outputs.

Another factor that we have briefly mentioned in this manuscript, but deserves further consideration is that of depth. Similar work exists, which measures depth in natural scenes (Calow & Lappe, 2007). Just as motion signal in the model varies with contrast it is also sensitive to how far away an object is. Distant buildings in one of the campus clips for example elicit noisy and small motion response as the further away something is, the less distance it will move in the clip. The movie clips used here are roughly equivalent in the distances of the objects in the scene, but this is not something we have controlled. In further studies, acquiring range data as well as motion data will be useful in disambiguating these two aspects.

In conclusion, by considering the statistical properties of local motion outputs we can find the limits on information available and the kind of differences the visual system has to overcome to extract self-motion. However these differences may not be totally discarded as they are useful for extracting object structure and finding differences between scenes. Combining different spatial scales can give important, more complete information as the location of motion signals in the scene varies over

different scales. Future work could further investigate the spatial variation in motion signals , extending this work and that of Roth & Black (2007), which considered the derivatives of the velocity field. From this work we conclude that large differences exist in the statistical properties of local motion caused my self-motion through natural scenes caused by the differential influence of the aperture effect. Although sparse local motion information does provide adequate information of extracting our heading direction and the large differences between different types of scenes may be crucial for fast gist perception during self-motion.

Acknowledgements

Thanks to Philip Roberts, Andrew Meso and Andrew Shaw. This work is supported The Leverhulme Trust Early Career Fellowship ECF/2007/0326

References

Barron, J. L., Fleet, D. J., Beachemin, S. S. (1994) Performance of optical flow techniques. *International Journal of Computer Vision* 12(1):43-77

Bartels A, Zeki S, Logothetis NK (2008) Natural vision reveals regional specialization to local motion and to contrast-invariant, global flow in the human brain. *Cerebral Cortex* 18, 705-717.

Calow, D., Krüger, N., Wörgötter, F., Lappe, M. (2004) Statistics of optic flow for self-motion through natural scenes. In *Dynamic Perception*, Ilg, U., Bülthoff, H. and Mallot, H. (eds.), pp. 133–138.

Calow, D., Lappe, M. (2007) Local statistics of retinal optic flow for self-motion through natural sceneries. *Network: Computation in neural systems.* 18(4): 343–374

Calow, D., Lappe, M. (2008) Efficient encoding in natural optic flow. *Network: Computation in neural systems.* 19(3)183-212

Coppola, D. M., Purves, H. R., McCoy, A. N., Purves, D. (1998) The distribution of oriented contours in the real world. *Proc. Natl. Acad. Sci. USA* 95:4002-4006

Cowey A., Rolls, E. T. (1974) Human cortical magnification factor and its relation to visual acuity. *Exp. Brain. Res.* 21:447-454

Dakin, S. C., Mareschal, I., Bex, P. J. (2005) An oblique effect for local motion: Psychophysics and natural movie statistics. *Journal of Vision* 5, 878-887

Durant, S., Johnston, A. (2006) Moving from segregated to transparent motion: a modelling approach. *Biology Letters* 2(1):101-105

Durant, S., Wall, M., B, Zanker, J. M. (2011) Manipulating the content of dynamic natural scenes to characterize response in human MT/MST. *Journal of Vision* 11, 10

Fleet, D. J., Langley, K.(1995) Recursive filters for optical flow. IEEE Trans Pattern *Analysis of Machine Intelligence* 17:61–67.

Friedman, A. (1979) Framing pictures: the role of knowledge in automatized encoding and memory for gist. 108(3):316-355

Gibson, JJ. 1950. *The perception of the visual world*. Cambridge, MA: The Riverside Press.

Hollands, M. A., Marple-Horvat D. E., Henkes, S., Rowan, A. K. (1995) Human eye movements during visually guided stepping. *Journal of Motor Behavior* 27(2):155-163

Horn, B. J. P., Schunk, B. G. (1981) Determining optical flow. *Artificial Intelligence* 17:185-203

Johnston, A., McOwan, P., Benton, C. (1999) Robust velocity computation from a biologically motivated model of motion perception. *Proceedings of the Royal Society: B* 250(1329):297-306

Lappe M. 2000. *Neuronal processing of optic flow*. San Diego: Academic Press

Lappe M., Bremmer, F., van den Berg, A. V. (1999) Perception of self-motion from visual flow. *Trends in Cognitive Sciences* 3(9):329-336

Lappin, J. S., Tadin, D., Nyquist, J. B., Corn, A. L. (2009) Spatial and temporal limits of motion perception across variations in speed, eccentricity, and low vision. *Journal of Vision* 9(1):30

McKee, S., Nakayama, K. (1984) The detection of motion in the visual peripheral field. *Vision Research* 24(1):25-32

Meso, A. I., Zanker, J. M. (2009) Speed encoding in correlation motion detectors as a consequence of spatial structure. *Biological Cybernetics* 100:361-370

Regan, D, Beverley, K. I. (1982) How do we avoid confounding the direction we are looking and the direction we are moving? *Science* 215: 194-196

Roth, S., Black, M. J. (2007) On the spatial statistics of optical flow. *Int. J. Comp. Vis.* 74(1):33-50

Simoncelli, E. P., Olshausen, B. A. (2001) Natural image statistics and neural representation. *Annual Reviews of Neuroscience* 24:1193-1216

Rousselet GA, Joubert OR, Fabre-Thorpe M (2005) How long to get to the "gist" of real-world natural scenes? *Visual Cognition* 12(6): 852-877

van der Schaaf, A., van Hateren, Y. H. (1996) Modelling the Power Spectra of Natural Images: Statistics and Information. *Vision Research* 36(17):2759-2770

Warren, Jr., W. H., Blackwell, A. W., Blackwell, K. J., Kurtz, N. G., Hatsopoulos, N. G., Kalish, M. L. (1991) On the sufficiency of the velocity field for perception of heading. *Biological Cybernetics* 65, 311-320

Warren, Jr., W. H., Morris, M. W., Kalish, M. (1988) Perception of translational heading from optical flow. *Journal of Experimental Psychology* 14(4):646-660

Wexler, W., Panerai, F., Lamouret, I., Droulez, J. (2001) Self-motion and the perception of stationary objects. *Nature* 409:85-88

Weiss, Y., Simoncelli, E. P., Adelson, E. H. (2002) Motion illusions as optimal percepts. *Nature Neuroscience* 5(6):598-604

Wiesel, T. (1982) Postnatal development of the visual cortex and the influence of environment. *Nature* 299(5884):583-591

Wilson, J. R., Sherman, S. M. (1976) Receptive-field characteristics of neurons in cat striate cortex - changes with visual-field eccentricity. *Journal of Neurophysiology* 39(3): 512-533

Zanker, J. M., Srinivasan, M. V., Egalhaaf, M. (1999) Speed tuning in elementary motion detectors of the correlation type. *Biological Cybernetics* 80:109-116

Zanker, J. M., Zeil, J. (2005) Movement-induced motion signal distributions in outdoor scenes. *Network: Comp. Neur. Syst.* 8(4):357-376

Zanker, J. M. (2001) Combining Local Motion Signals: A Computational Study of Segmentation and Transparency. In: Zanker, & Zeil (Eds.), *Motion Vision: - Computational, Neural, and Ecological Constraints* (pp. 113-124). Berlin Heidelberg New York: Springer.